

Do AI weather models forecast extremes better than GFS? A station-based verification over China with a common initialization source, 2022-2025

Yi Zou

yizou505@gmail.com

Independent Researcher <https://orcid.org/0009-0000-6493-7255>

Research Article

Keywords: AI weather prediction, extreme weather, forecast verification, potential CRPS, threshold-weighted CRPS, initial conditions

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11259821/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Do AI weather models forecast extremes better
than GFS? A station-based verification over China
with a common initialization source, 2022–2025

Yi Zou^{1*}

¹Independent researcher, Xi'an, China.

Corresponding author(s). E-mail(s): yizou505@gmail.com;

Abstract

Whether artificial intelligence (AI) weather models are more informative for extremes depends on verification and initialization. We compare GraphCast Operational, Pangu-Weather and FourCastNet v2-small with the operational Global Forecast System (GFS) at 334 surface stations in China for 00 UTC initializations from January 2022 to August 2025. The AI runs share the GFS F000 input source, controlling initialization source while retaining model-specific training and preprocessing. Potential continuous ranked probability scores (PCRPS) and threshold-weighted versions assess information after in-sample isotonic distributional regression (IDR). Independent-period IDR and Gaussian ensemble model output statistics (EMOS) assess achievable post-processing skill. The temperature comparison favours the AI models at the reported daily leads, with exceptions at six-hourly leads. GraphCast's heavy-precipitation advantage weakens out of sample, where tail-skill intervals include zero beyond day 1. Temperature rankings differ between verification settings, skill varies across climate groups, and initialization from the Integrated Forecasting System (IFS) substantially changes short-range gains. Comparisons of extremes should distinguish potential information, achievable skill and initialization sensitivity.

Keywords: AI weather prediction, extreme weather, forecast verification, potential CRPS, threshold-weighted CRPS, initial conditions

1 Introduction

Data-driven weather models trained on reanalysis, such as GraphCast (Lam et al. 2023), Pangu-Weather (Bi et al. 2023) and the spherical Fourier neural operator (SFNO) behind FourCastNet v2-small (Bonev et al. 2023), now match or exceed operational deterministic forecasts on many headline scores (Rasp et al. 2024; Ben Bouallègue et al. 2024). Whether this also holds for extreme events, the cases that matter most for warnings, is disputed. Olivetti and Messori (2024) found no systematic weakness of AI models for extremes on WeatherBench 2 fields. In contrast, Zhang et al. (2026) reported that physics-based models outperform AI forecasts of record-breaking temperature and wind extremes. Conversely, Biegert et al. (2026) found that AI models, FuXi in particular, are *more* informative for extremes when they are compared with the threshold-weighted potential CRPS against a common reanalysis truth. They attributed the disagreement to the forecaster’s dilemma (Lerch et al. 2017): conditioning the evaluation on observed extremes favours forecasters who predict extremes more often. They also noted that Zhang et al. (2026) verified each model against a different ground truth.

Three design choices underlie these conflicting results. The first is the *verification metric*. Point-error metrics such as the root mean square error (RMSE) reward the smooth, mean-like forecasts that AI models are trained to produce. Gneiting et al. (2026) therefore proposed the potential CRPS (PCRPS). It converts each deterministic forecast into a calibrated predictive distribution by isotonic distributional regression (IDR, or EasyUQ; Henzi et al. 2021; Walz et al. 2024) and then scores that distribution. Threshold-weighted versions of the score focus on the tails without the forecaster’s dilemma (Gneiting and Ranjan 2011; Allen et al. 2023; Biegert et al. 2026).

The second choice is the *ground truth*. Gridded (re)analyses are smooth, are produced by a particular forecast system, and are often the training target of the AI

models. Station observations provide a measurement-based verification target but include point-to-grid representativeness error and uneven sampling. Related regional work includes temperature verification (He et al. 2026) and GraphCast precipitation verification at 2393 Chinese stations over 2020–2021 (Yan et al. 2025). Our contribution is the combination of a common initialization source, multiple variables and years, climate-group comparisons, and separate in-sample and out-of-sample tail verification; the scoring methods themselves are established.

The third choice is the *initial conditions*. Most published comparisons pit AI models initialized from ERA5 or IFS analyses against a physics model with its own analysis (Rasp et al. 2024; Ben Bouallègue et al. 2024; Charlton-Perez et al. 2024). Differences in skill then mix model and analysis quality. The NOAA/CIRA AI weather prediction (AIWP) archive (Radford et al. 2025) permits comparisons conditional on a shared initialization source, and runs each AI system with both GFS and IFS inputs. Differences can still involve preprocessing and interactions between training and initialization.

Here, we combine all three elements. We verify GFS-initialized GraphCast Operational, Pangu-Weather and FourCastNet v2-small against operational GFS forecasts, controlling the source of the initial fields. The verification uses observations from 334 stations in China between 2022 and 2025. China spans tropical to highland and arid climates and experiences frequent heat waves, cold surges and heavy rainfall. We evaluate 2-m temperature, daily maximum and minimum temperature, 10-m wind speed and 24-h precipitation. Each variable is scored with the PCRPS and with twPCRPS focused on station-specific climatological tails. These in-sample results are complemented by out-of-sample post-processing, block-bootstrap confidence intervals and multiplicity-controlled station-wise tests. Finally, we repeat the analysis with the IFS-initialized versions of the same models. This measures initialization sensitivity conditional on each trained AI system, rather than an additive causal decomposition of architecture and analysis quality.

2 Data

2.1 Forecasts

AI forecasts were taken from the NOAA/CIRA AIWP archive (Radford et al. 2025), which is hosted on Amazon Web Services (bucket `noaa-oar-mlwp-data`). We used GraphCast Operational (archive identifier `GRAP_v100`), Pangu-Weather (`PANG_v100`) and FourCastNet v2-small (`FOUR_v200`), with both GFS and IFS initialization. The operational GraphCast checkpoint was fine-tuned on IFS data through 2021. The GFS-initialized archive uses the 0.25° F000 products rather than the separate analysis products (Radford et al. 2025). Archive identifiers and initialization source are held fixed within each comparison; training, variable preprocessing and learned dynamics remain system-specific. All models provide 0.25° global fields every 6 h. The physics-based reference is the operational deterministic GFS at 0.25° , retrieved from the NOAA GFS archive on AWS (`noaa-gfs-bdp-pds`) using byte ranges from the GRIB index files. We considered 00 UTC initializations from 1 January 2021 to 12 August 2025 and leads of 6–240 h. Forecasts were bilinearly interpolated to the station locations.

The archive has gaps. GFS-initialized GraphCast is unavailable before 31 December 2021, and five initializations in 2024 and two in 2025 are missing. Pangu-Weather and FourCastNet v2-small each miss one 2021 initialization. The IFS-initialized runs miss a few dates in 2022–2023 and contain longer gaps in 2024. One GFS initialization (2 February 2021) could not be retrieved. Because GraphCast starts at the end of 2021, the main comparison begins in January 2022 (Section 3.1). GraphCast precipitation, archived as 6-h accumulations in metres, was converted to millimetres. Among the AI models, only GraphCast provides precipitation.

2.2 Station observations

Observations come from the Integrated Surface Database (ISD; Smith et al. 2011). Starting from 360 candidate stations in China (WMO index blocks 50–59), we retained stations that reported at least 80% of 00 UTC observations between 2021 and August 2025. They also had to have at least 80% coverage in at least 8 of the 10 years 2011–2020, which form the climatological reference period. This yielded 334 stations (Fig. 1). Two-metre temperature and 10-m wind speed were taken at the synoptic hours from FM-12 SYNOP reports within ± 10 min of the hour, where available.

Daily maximum and minimum temperature ($T_{\max}$, $T_{\min}$) were taken from the ISD extreme-temperature groups, which refer to the 24 h ending at 12 UTC. These groups are only available from late 2019, and their coverage drops in 2024–2025. Gaps were therefore filled with quality-screened GHCN-Daily TMAX/TMIN (Menne et al. 2012), matched by station and reported date. The median absolute differences on overlapping 2020–2023 days are 0.0 °C for $T_{\max}$ and 0.1 °C for $T_{\min}$. This agreement supports the date matching, but does not verify identical observation windows for every source record. After filling, 180 stations meet the completeness requirements for $T_{\max}$ and $T_{\min}$. Daily precipitation is the 24-h accumulation reported at 00 UTC (00–00 UTC window). ISD AA amount code 9999 is missing, and condition code 9 is not a zero-precipitation flag (NOAA National Centers for Environmental Information 2018). A missing AA 24-h amount is recovered as zero only when a valid FM-12 source report explicitly contains the unambiguous 70000 group in SYNOP section 333 (World Meteorological Organization 1995, regulation 12.4.9). Numeric AA totals with accumulation-boundary flags require agreement with the original SYNOP 24-h group. Other incomplete, suspect, unresolved and conflicting records remain missing. Parser versions, source hashes and recovery counts are recorded per station-year. GHCN-Daily comparisons use quality-screened values, retain the daily-window difference, and

do not establish that an ISD missing code means zero. In total, 249 stations are used for precipitation.

Stations were assigned to the main climate classes of the 0.1° Köppen–Geiger map of Beck et al. (2023). Stations in polar (E) classes or above 3000 m elevation were grouped as *Highland*. This yields 149 Temperate, 76 Arid, 75 Cold, 28 Highland and 6 Tropical stations. For $T_{\max}/T_{\min}$ the corresponding counts are 64, 54, 41, 18 and 3.

2.3 Derived forecast variables

Forecast daily $T_{\max}$ and $T_{\min}$ for day k ($k = 1, \dots, 9$) are the maximum and minimum of the four 6-hourly 2-m temperature values at 18, 00, 06 and 12 UTC, i.e. leads $24k - 6$ to $24k + 12$ h. They match the observed 12–12 UTC window. The same sampling is applied to all models, including GFS. This controls the sampling convention but does not ensure identical errors, because the timing and amplitude of each model’s diurnal cycle can differ. Daily precipitation for day k ($k = 1, \dots, 10$) is the sum of the four 6-h accumulations ending at leads $24(k - 1) + 6, \dots, 24k$ h.

2.4 Extreme thresholds

Tail thresholds are station-specific quantiles of the 2011–2020 climatology (Table 1). For 2-m temperature and wind speed, they are also specific to the valid hour. For $T_{\max}/T_{\min}$, the climatology combines ISD and GHCN-Daily. The completeness screen for both daily-temperature thresholds uses the $T_{\max}$ reference series, requiring at least 1000 days and at least 50 days in every calendar month; each variable’s quantiles are then computed from its available values. For precipitation, ISD was supplemented with GHCN-Daily where ISD provided fewer than 1500 days. This mixes 00–00 and 12–12 UTC climatological totals at those stations and may affect wet-day quantiles; the source is recorded in the climatology output and should be considered when interpreting tail skill.

Table 1 Lower or upper tails used for the threshold-weighted scores. Quantiles q_p are station-specific and computed from 2011–2020 observations; for 2-m temperature and wind speed they are also specific to the valid hour. The wet-day quantile uses days with at least 0.1 mm.

Variable	Tail 1 (tw ₁)	Tail 2 (tw ₂)
2-m temperature	$y > q_{0.95}$	$y < q_{0.05}$
Daily $T_{\max}$	$y > q_{0.90}$	$y > q_{0.95}$
Daily $T_{\min}$	$y < q_{0.10}$	$y < q_{0.05}$
10-m wind speed	$y > q_{0.90}$	$y > q_{0.95}$
24-h precipitation	$y > \text{wet-day } q_{0.90}$	$y \geq 25 \text{ mm}$

3 Methods

Processing scripts, station lists, derived tables and provenance records are publicly available in the repository identified in the Data and code availability statement. Supplementary key-result tables and the station inventory are provided in Online Resource 1 (ESM_1.pdf).

3.1 Comparison sets

All scores are computed on a common sample: for a given station, variable and lead time, only cases in which every compared model and the observation are available are used. The main comparison (*core*) includes GFS and the three GFS-initialized AI models for initializations from January 2022 to August 2025. For 2-m temperature it contains on average about 1225 cases per station and lead. Where the bootstrap is computed from daily station sums, about 1300 distinct days enter. A sensitivity set (*core21*) adds 2021 but omits GraphCast. The *ic* set adds the three IFS-initialized models and is restricted to their common initializations (about 1050 cases per station and lead for 2-m temperature).

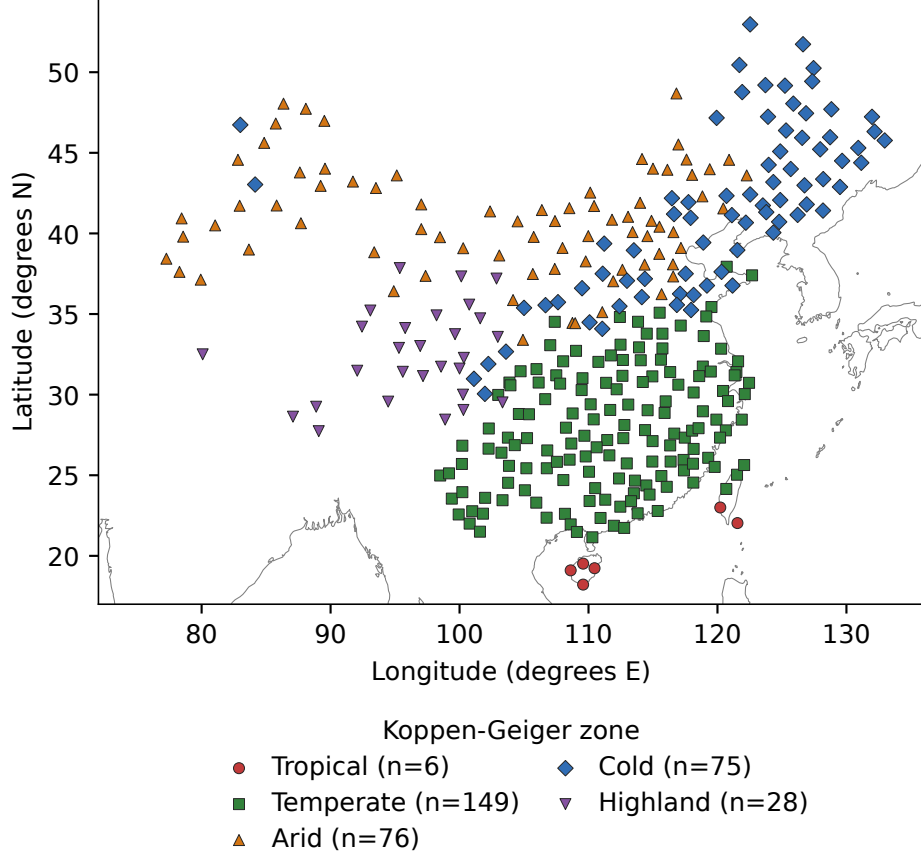


Fig. 1 The 334 ISD stations used in this study, classified by Köppen-Geiger climate group (Beck et al. 2023). Marker shapes identify climate groups in monochrome printing.

145 3.2 Potential CRPS and threshold-weighted PCRPS

For a predictive cumulative distribution function (CDF) F and observation y , the continuous ranked probability score is (Matheson and Winkler 1976; Gneiting and Raftery 2007)

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{I}\{y \leq z\})^2 dz. \quad (1)$$

The threshold-weighted CRPS with indicator weight $w(z) = \mathbb{I}\{z > t\}$ (or $\mathbb{I}\{z < t\}$) restricts the integral to the tail of interest (Gneiting and Ranjan 2011; Allen et al.

2023):

$$\text{twCRPS}_t(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{I}\{y \leq z\})^2 w(z) \, dz. \quad (2)$$

Following Gneiting et al. (2026), the deterministic forecasts $x_1, \dots, x_n$ of a model at one station and lead time are converted into predictive distributions $\hat{F}_1, \dots, \hat{F}_n$ by IDR fitted to the pairs (x_i, y_i) (Henzi et al. 2021; Walz et al. 2024). IDR yields the CRPS-optimal distributions under the sole constraint that a larger forecast implies a stochastically larger outcome. Since IDR is simultaneously optimal for the threshold-weighted CRPS (Henzi et al. 2021), the mean scores

$$\text{PCRPS} = \frac{1}{n} \sum_{i=1}^n \text{CRPS}(\hat{F}_i, y_i), \quad \text{twPCRPS}_t = \frac{1}{n} \sum_{i=1}^n \text{twCRPS}_t(\hat{F}_i, y_i) \quad (3)$$

146 measure the information content of the deterministic forecast. These scores are invari-
 147 ant to strictly increasing transformations of the deterministic forecasts and quantify
 148 discrimination after the monotone recalibration allowed by IDR. Smoothing that
 149 changes forecast ordering or creates ties can still affect them (Gneiting et al. 2026;
 150 Biegert et al. 2026). All twPCRPS values use the thresholds of Table 1. Our imple-
 151 mentation was checked against brute-force numerical quadrature (agreement better
 152 than 10^{-4}).

153 Skill relative to GFS is reported as $\text{SS} = 1 - \sum S_{\text{model}} / \sum S_{\text{GFS}}$, where the sums
 154 run over all stations and cases. Positive values mean that the AI model is better. We
 155 also computed raw bias, mean absolute error and RMSE at each station. The summary
 156 CSV reports equal-station means of those station statistics; the skill estimates and
 157 bootstrap tables use pooled station-case score sums. Mean station RMSE is not pooled
 158 RMSE. These station-sampled aggregates are not area-weighted estimates for China.

159 3.3 Out-of-sample post-processing

160 The PCRPS is an in-sample measure. To test whether potential skill transfers to post-
161 processed forecasts, we also fitted two methods on 2022–2023 (2021–2023 for *core21*)
162 and evaluated them on January 2024–August 2025. The calibration is frozen before the
163 first test initialization, 1 January 2024. For each variable and lead, training cases are
164 purged unless their target observation window ends strictly before that time, including
165 the extra 12 h in daily temperature targets. This prevents using training labels that
166 would not yet have been observed. Reporting delays beyond the observation-window
167 end are not modelled. The first method is out-of-sample IDR. Predictive CDFs for
168 new forecasts are linearly interpolated between the CDFs of the neighbouring training
169 forecasts. The second is Gaussian ensemble model output statistics (EMOS; Gneiting
170 et al. 2005), applied to 2-m temperature (at 24-h multiples of lead time), $T_{\max}$ and
171 $T_{\min}$. Its predictive mean is $a + bx$ plus first-harmonic seasonal terms, and the log-
172 arithm of its standard deviation has a seasonal cycle. Parameters are estimated by
173 minimum-CRPS fitting.

174 3.4 Statistical inference

175 Aggregate skill uncertainty was estimated with a moving-block bootstrap (Kunsch
176 1989) with 14-day blocks and 1000 replicates. The bootstrap resamples paired 14-
177 day blocks on the initialization-date calendar, using station-summed losses. Valid zero
178 losses are retained, including days on which the GFS loss is zero. Missing days con-
179 tribute zero to both sums with availability recorded separately and keep their calendar
180 positions. All stations share the same sampled blocks, preserving the observed within-
181 day spatial dependence and the dependence represented within each block. Percentile
182 95% intervals are reported when at least 95% of replicates have a positive reference
183 total. These intervals condition on fitted in-sample IDR maps; they do not quantify
184 all calibration-estimation uncertainty. At each station, we tested the null hypothesis

185 of equal predictive performance with the Diebold–Mariano test (Diebold and Mar-
 186 iano 1995). Its variance was estimated by the Newey–West heteroskedasticity- and
 187 autocorrelation-consistent (HAC) estimator (Newey and West 1987), with lag trunca-
 188 tion $\max(\text{lead in days}, n^{1/3})$. Missing dates retain their calendar spacing in the HAC
 189 calculation. The resulting p values were adjusted across stations by the Benjamini–
 190 Hochberg procedure (Benjamini and Hochberg 1995) at a false discovery rate of 5%
 191 separately for each model, reference, variable, lead and score, following Wilks (2016).
 192 We report the percentage of stations at which an AI model is significantly better
 193 or worse than GFS. The bootstrap was also computed separately for each climate
 194 group. This multiplicity adjustment is not a study-wide correction over all scores and
 195 leads; spatial dependence and fitted calibration maps also qualify the interpretation
 196 of station-wise tests. All six-hourly leads are saved for instantaneous variables, with
 197 EMOS restricted to 24-h multiples.

198 **3.5 Sensitivity to the initialization source**

199 In the *ic* set, each architecture is available from both GFS and IFS initial conditions.
 200 The GFS-initialized comparison controls the source of the initial fields and measures
 201 differences between forecast systems conditional on that source. The paired change
 202 after switching to IFS measures initialization sensitivity for the fixed trained AI sys-
 203 tem. It includes compatibility with the model’s training data and preprocessing, so it
 204 cannot identify a pure architecture effect or an independent additive contribution of
 205 analysis quality.

206 **3.6 Use of AI tools**

207 GPT-5.6 (OpenAI) was used solely to polish the language of this manuscript. The
 208 author reviewed and edited the resulting text and takes full responsibility for the
 209 content of the manuscript.

210 4 Results

211 Tables 2 and 3 summarise the skill of the three AI models relative to GFS in the *core*
212 set. Figure 2 shows the lead-time dependence available in the result files. Tables report
213 24-h multiples for instantaneous variables; skill at every six-hourly lead, separated by
214 valid hour, is shown in Fig. 3.

215 4.1 Two-metre temperature

216 All three AI models show positive PCRPS skill for 2-m temperature at the reported
217 24-h multiples (Table 2). GraphCast improves the PCRPS by 8.0% at 24 h, 11.0% at
218 120 h and 8.5% at 240 h. It is significantly better than GFS at 70–83% of the stations,
219 depending on lead, and significantly worse at no more than 11%. Pangu-Weather has
220 the lowest PCRPS point estimates at the reported daily leads up to about 5 days (8.4%
221 at 24 h, 11.5% at 120 h), but its advantage decays to 3.7% at 240 h. FourCastNet v2-
222 small improves by 5–7% at those leads. This does not imply superiority at every
223 six-hourly lead: at 6 h its PCRPS skill is -1.6% , with a bootstrap interval excluding
224 zero (Fig. 3). The equal-station mean PCRPS at 24 h is 0.957 K for GFS and 0.877–
225 0.900 K for the AI models. The raw forecasts show the same ordering. At 24 h, the
226 equal-station mean RMSE is 2.61 K for GFS and 2.38 K for Pangu-Weather; GFS has
227 a mean warm bias of $+0.34$ K, whereas GraphCast has a cold bias of -0.22 K.

228 The advantage persists for warm extremes. The twPCRPS for exceedances of the
229 95th percentile shows skill of 6–12% at the reported 24-h multiples. For cold extremes
230 (below the 5th percentile), GraphCast and Pangu-Weather are significantly better
231 than GFS up to 120 h. FourCastNet v2-small loses skill with lead time; at 240 h its
232 cold-tail skill is -6.0% , but the interval includes zero.

Table 2 Skill (%) of the GFS-initialized AI models relative to GFS for the temperature variables in the *core* set (initializations January 2022–August 2025). Positive values indicate that the AI model is better. Bold: 95% block-bootstrap confidence interval excludes zero. Brackets (PCRPS only): percentage of stations at which the model is significantly better/worse than GFS (Diebold–Mariano test, Benjamini–Hochberg false discovery rate 5%). tw_1 , tw_2 : threshold-weighted PCRPS for the tails in Table 1. OOS IDR: out-of-sample IDR trained on 2022–2023 and tested on January 2024–August 2025.

Score	Model	Lead time		
<i>2-m temperature</i> (leads: 24 h, 120 h, 240 h)				
PCRPS	GraphCast Op.	+8.0 [70/11]	+11.0 [81/2]	+8.5 [78/1]
PCRPS	Pangu-Weather	+8.4 [69/11]	+11.5 [79/3]	+3.7 [25/1]
PCRPS	FCN v2-small	+5.9 [56/20]	+7.1 [57/9]	+5.1 [44/1]
twPCRPS ₁	GraphCast Op.	+9.4	+9.1	+8.0
twPCRPS ₁	Pangu-Weather	+8.9	+11.2	+6.4
twPCRPS ₁	FCN v2-small	+6.9	+8.1	+6.0
twPCRPS ₂	GraphCast Op.	+5.1	+5.0	+1.8
twPCRPS ₂	Pangu-Weather	+7.2	+7.9	+0.7
twPCRPS ₂	FCN v2-small	+3.7	−3.1	−6.0
OOS IDR	GraphCast Op.	+7.4	+10.2	+7.6
OOS IDR	Pangu-Weather	+8.1	+11.4	+3.4
OOS IDR	FCN v2-small	+5.2	+6.3	+3.4
<i>Daily T_{max}</i> (leads: d1, d5, d9)				
PCRPS	GraphCast Op.	+8.0 [62/18]	+9.2 [68/5]	+7.1 [61/1]
PCRPS	Pangu-Weather	+6.1 [43/20]	+7.4 [48/5]	+0.6 [26/16]
PCRPS	FCN v2-small	+2.7 [34/22]	+5.6 [36/4]	+4.1 [33/6]
twPCRPS ₁	GraphCast Op.	+2.8	+5.5	+7.5
twPCRPS ₁	Pangu-Weather	+4.3	+6.3	+4.9
twPCRPS ₁	FCN v2-small	+1.7	+5.4	+4.8
twPCRPS ₂	GraphCast Op.	+0.3	+2.5	+6.0
twPCRPS ₂	Pangu-Weather	+1.5	+3.3	+3.5
twPCRPS ₂	FCN v2-small	−1.7	+2.7	+3.1
OOS IDR	GraphCast Op.	+6.4	+7.6	+6.6
OOS IDR	Pangu-Weather	+5.7	+6.6	+0.6
OOS IDR	FCN v2-small	+3.6	+4.5	+1.6
<i>Daily T_{min}</i> (leads: d1, d5, d9)				
PCRPS	GraphCast Op.	+8.9 [75/8]	+11.7 [83/0]	+8.4 [71/0]
PCRPS	Pangu-Weather	+14.9 [88/6]	+13.6 [86/2]	+4.8 [38/1]
PCRPS	FCN v2-small	+11.1 [75/9]	+8.9 [62/5]	+4.5 [46/2]
twPCRPS ₁	GraphCast Op.	+4.8	+6.9	+3.2
twPCRPS ₁	Pangu-Weather	+9.7	+10.0	+2.1
twPCRPS ₁	FCN v2-small	+6.3	+0.5	−5.2
twPCRPS ₂	GraphCast Op.	+4.7	+5.8	+3.7
twPCRPS ₂	Pangu-Weather	+9.3	+8.1	+0.9
twPCRPS ₂	FCN v2-small	+5.6	−1.9	−7.0
OOS IDR	GraphCast Op.	+6.3	+8.7	+5.9
OOS IDR	Pangu-Weather	+10.1	+10.5	+2.9
OOS IDR	FCN v2-small	+6.3	+6.4	+1.8

Table 3 As Table 2, but for 10-m wind speed and 24-h precipitation (GraphCast only).

Score	Model	Lead time		
<i>10-m wind speed</i> (leads: 24 h, 120 h, 240 h)				
PCRPS	GraphCast Op.	+0.6 [19/15]	+1.8 [15/3]	+0.8 [6/1]
PCRPS	Pangu-Weather	-1.7 [11/29]	+1.3 [14/7]	+0.1 [1/2]
PCRPS	FCN v2-small	-6.4 [5/60]	-1.8 [6/24]	+0.0 [3/3]
twPCRPS ₁	GraphCast Op.	+1.0	+2.4	+1.2
twPCRPS ₁	Pangu-Weather	-2.4	+1.6	+0.6
twPCRPS ₁	FCN v2-small	-7.4	-1.2	+0.4
twPCRPS ₂	GraphCast Op.	+0.7	+2.4	+1.4
twPCRPS ₂	Pangu-Weather	-2.5	+1.4	+0.9
twPCRPS ₂	FCN v2-small	-7.4	-1.2	+0.5
OOS IDR	GraphCast Op.	+0.9	+1.6	+0.6
OOS IDR	Pangu-Weather	-1.1	+1.6	-0.2
OOS IDR	FCN v2-small	-5.3	-1.2	+0.0
<i>24-h precipitation</i> (leads: d1, d5, d10)				
PCRPS	GraphCast Op.	+4.7 [8/1]	+6.2 [14/0]	+2.6 [0/0]
twPCRPS ₁	GraphCast Op.	+1.5	+2.7	+1.0
twPCRPS ₂	GraphCast Op.	+1.7	+3.3	+1.2
OOS IDR	GraphCast Op.	+4.3	+3.6	+1.0

4.2 Daily maximum and minimum temperature

GraphCast has the most consistent advantage for $T_{\max}$ (7–10% PCRPS skill over days 1–9). Pangu-Weather’s skill falls from 6–8% at days 1–5 to an insignificant 0.6% at day 9, and FourCastNet v2-small gains 3–6%. For the warm tail (above the 90th percentile), the skill of GraphCast increases with lead time, from 2.8% at day 1 to 7.5% at day 9. At days 1–5, all four models underestimate $T_{\max}$ by about 2–3.5 °C on average (not shown in the table). This is expected, because the six-hourly samples miss the afternoon peak. PCRPS removes a constant or strictly monotone recalibration error, but six-hourly sampling can also change forecast ordering and discrimination; those effects are not removed by calibration.

Improvements are largest for $T_{\min}$. Pangu-Weather reaches 14.9% skill at day 1 and 13.6% at day 5, and is significantly better than GFS at 86–88% of stations up to day 5. In the cold tail (below the 10th percentile), Pangu-Weather and GraphCast retain significant skill up to day 7. FourCastNet v2-small becomes significantly worse

247 than GFS at day 9 (-5.2%). At days 1–5, GFS overestimates $T_{\min}$ by 1.3–1.4 °C on
248 average, more than any of the AI models.

249 4.3 Ten-metre wind speed

250 Differences for wind speed are small (Table 3). GraphCast has small positive skill
251 at the reported 24-h multiples (0.6–1.8%). At 6 h, however, its skill is significantly
252 negative (-1.3% ; Fig. 3), so a claim of superiority at all leads is unsupported. Pangu-
253 Weather is worse at 24 h (-1.7%) and slightly better at 3–7 days. FourCastNet v2-
254 small is clearly worse at short leads (-6.4% at 24 h, significantly worse at 60% of
255 stations), and its deficit vanishes by 240 h. The rankings for the strong-wind tails
256 mirror those for the full distribution.

257 4.4 Precipitation

258 GraphCast improves the PCRPS of 24-h precipitation by 4.7% on day 1, 6.2–6.5% on
259 days 3–5 and 2.6% on day 10; all improvements are significant (Table 3). The skill
260 for heavy precipitation is smaller but significant at all leads. For exceedances of the
261 wet-day 90th percentile it is 1.0–3.7%, and for the absolute threshold of 25 mm it is
262 1.2–4.1%. Station-wise tests rarely reach significance for precipitation (at most 14%
263 of stations; Fig. 4), reflecting its large sampling variability.

264 4.5 Climate zones

265 The AI advantage is largest in the Temperate zone, which contains most stations in
266 eastern China, and is weaker in the Arid and Highland zones (Table 4, Fig. 5). For 2-m
267 temperature at 120 h, GraphCast and Pangu-Weather improve the PCRPS by 13.8%
268 and 16.7% at Temperate stations. At Highland stations, GraphCast’s skill falls to an
269 insignificant 2.5%, while Pangu-Weather retains 13.6%. For $T_{\max}$ at day 5, GraphCast
270 is significantly *worse* than GFS at Highland stations (-7.3%), while all models gain
271 14–18% at Temperate stations. For $T_{\max}$ at Arid stations, the bootstrap intervals

for Pangu-Weather and FourCastNet v2-small include zero, so these comparisons do not establish a difference from GFS. For precipitation, GraphCast gains 8.8% in the Temperate zone at day 5, compared with 1.5% in the Arid zone. For wind speed, the skill is close to zero or slightly negative in the Cold zone. The Tropical group comprises only six stations (three for $T_{\max}/T_{\min}$), so its large values should be interpreted with caution.

Table 4 Skill (%) relative to GFS by climate group at a medium lead time (*core* set). Bold: 95% block-bootstrap confidence interval excludes zero. Station counts for instantaneous variables ($T_{\max}/T_{\min}$ in parentheses): Temperate 149 (64), Arid 76 (54), Cold 75 (41), Highland 28 (18), Tropical 6 (3). Precipitation counts are 124, 47, 50, 24 and 4 respectively (249 total).

Score	Model	All	Temperate	Arid	Cold	Highland	Tropical
<i>2-m temperature, 120 h</i>							
PCRPS	GraphCast Op.	+11.0	+13.8	+10.0	+10.8	+2.5	+16.6
PCRPS	Pangu-Weather	+11.5	+16.7	+7.8	+7.4	+13.6	+14.5
PCRPS	FCN v2-small	+7.1	+10.1	+4.8	+5.0	+8.4	+0.7
twPCRPS ₁	GraphCast Op.	+9.1	+11.3	+7.4	+7.5	+9.7	+18.9
twPCRPS ₁	Pangu-Weather	+11.2	+15.3	+7.3	+9.3	+12.1	+10.7
twPCRPS ₁	FCN v2-small	+8.1	+11.3	+4.8	+7.9	+5.2	-1.6
<i>Daily $T_{\max}$, day 5</i>							
PCRPS	GraphCast Op.	+9.2	+17.2	+4.4	+9.0	-7.3	+21.1
PCRPS	Pangu-Weather	+7.4	+17.5	-1.6	+3.0	+6.6	+20.0
PCRPS	FCN v2-small	+5.6	+14.4	-1.3	+2.8	+0.3	+12.9
twPCRPS ₁	GraphCast Op.	+5.5	+8.5	+2.7	+5.5	+4.7	+2.1
twPCRPS ₁	Pangu-Weather	+6.3	+11.4	-1.2	+5.9	+14.3	+0.1
twPCRPS ₁	FCN v2-small	+5.4	+8.7	+0.3	+6.1	+10.3	-7.1
<i>Daily $T_{\min}$, day 5</i>							
PCRPS	GraphCast Op.	+11.7	+15.8	+11.1	+10.8	+5.2	+12.1
PCRPS	Pangu-Weather	+13.6	+17.7	+11.8	+10.5	+16.4	+10.7
PCRPS	FCN v2-small	+8.9	+12.1	+7.9	+5.4	+12.4	-1.2
twPCRPS ₁	GraphCast Op.	+6.9	+12.4	+5.0	+1.2	+7.9	+11.7
twPCRPS ₁	Pangu-Weather	+10.0	+13.9	+6.2	+4.2	+18.3	+20.1
twPCRPS ₁	FCN v2-small	+0.5	+4.4	-4.5	-6.5	+14.0	+7.1
<i>10-m wind speed, 120 h</i>							
PCRPS	GraphCast Op.	+1.8	+3.1	+1.6	-0.7	+0.9	+7.5
PCRPS	Pangu-Weather	+1.3	+3.0	+0.6	-1.0	+0.8	+6.1
PCRPS	FCN v2-small	-1.8	-0.9	-2.5	-3.4	-0.7	+3.2
twPCRPS ₁	GraphCast Op.	+2.4	+3.3	+2.8	+0.7	+2.0	+2.4
twPCRPS ₁	Pangu-Weather	+1.6	+3.1	+0.0	-0.1	+0.9	+4.7
twPCRPS ₁	FCN v2-small	-1.2	-0.5	-2.0	-1.9	-1.9	+0.5
<i>24-h precipitation, day 5</i>							
PCRPS	GraphCast Op.	+6.2	+8.8	+1.5	+2.4	+6.0	+17.8
twPCRPS ₁	GraphCast Op.	+2.7	+4.7	+0.8	+1.1	+0.0	+18.7

4.6 Out-of-sample post-processing

Temperature models generally retain positive overall skill in the independent 2024–2025 test period, but neither model rankings nor tail improvements transfer uniformly (Table 2, Fig. 6). With out-of-sample IDR, GraphCast’s skill for 2-m temperature is 7.4%, 10.2% and 7.6% at 24, 120 and 240 h, close to its in-sample values. Gaussian EMOS gives similar results, with skill of 7.1%, 8.4% and 4.7% for GraphCast and 7.3%, 10.9% and 4.2% for Pangu-Weather. For precipitation, out-of-sample IDR retains 1.0–4.3% overall skill (significant except at day 10), while the wet-day-q90 tail advantage is much smaller (Table 5). Only the day-1 tail skill is significant (+1.75%); the intervals include zero at the other nine leads, and the point estimates at days 3, 5 and 7 are 1.1%, 0.4% and 0.1%. In sample, the same tail skill is 1.0–3.7% and significant at every lead. For the temperature tails, out-of-sample EMOS twCRPS skill is positive and significant at all reported leads for warm 2-m temperature and $T_{\max}$ extremes, whereas cold-tail $T_{\min}$ skill is less consistent (FourCastNet v2-small: -6.1% at day 9). Ranking changes also occur for temperature: at day 1 for $T_{\max}$, in-sample PCRPS favours GraphCast over Pangu-Weather (8.0% versus 6.1%), whereas out-of-sample EMOS favours Pangu-Weather (7.8% versus 6.3%). This is a reversal of point estimates, not a paired significance test between the AI models. The in-sample and out-of-sample comparisons also use different evaluation periods, so this contrast does not isolate the effect of calibration alone. The *core21* sensitivity set, which uses three years of training, gives consistent results (e.g. Pangu-Weather PCRPS skill for 2-m temperature of 8.5%, 11.5% and 2.8% at 24, 120 and 240 h).

4.7 Sensitivity to the initialization source

Table 6 and Fig. 7 compare each trained AI system started from GFS and from IFS F000 inputs, on the common *ic* sample. Replacing the initial conditions more than doubles the skill of GraphCast for 2-m temperature at 24 h, from 8.4% to 18.8%

Table 5 GraphCast skill (%) for the wet-day-q90 precipitation tail: in-sample twPCRPS versus independent-period IDR twCRPS, with 95% bootstrap intervals.

Day	In-sample	95% interval	OOS IDR	95% interval
1	+1.52	[+0.69, +2.49]	+1.75	[+0.12, +4.15]
2	+2.99	[+1.91, +4.64]	+1.09	[-0.70, +3.53]
3	+3.70	[+2.36, +5.46]	+1.10	[-0.31, +2.74]
4	+3.67	[+2.41, +5.41]	+0.36	[-0.71, +1.62]
5	+2.71	[+1.35, +4.43]	+0.45	[-0.79, +1.67]
6	+2.93	[+1.72, +4.61]	+0.51	[-0.63, +2.06]
7	+2.75	[+1.53, +4.36]	+0.06	[-0.72, +1.17]
8	+1.90	[+0.95, +3.23]	-0.62	[-1.62, +0.18]
9	+1.88	[+1.04, +2.94]	+0.43	[-0.25, +1.18]
10	+0.96	[+0.18, +1.83]	+1.21	[-0.18, +2.59]

relative to GFS. The skill for $T_{\max}$ at day 1 rises from 7.9% to 19.0%, and for day-5 precipitation from 6.2% to 11.7%. For Pangu-Weather, the 24-h temperature skill rises from 8.5% to 17.1%. For FourCastNet v2-small the increase is smaller (6.0% to 8.1%). The short-range skill differences are large relative to the GFS-initialized gains, demonstrating sensitivity to the initialization source. They do not identify a causal fraction attributable solely to analysis quality, because the trained models and initialization data interact. The GFS-initialized results remain the relevant conditional comparison for a system supplied with GFS fields.

Table 6 PCRPS skill (%) relative to GFS of each trained AI system initialized from GFS $\rightarrow$ from IFS F000 fields, on the common sample of the *ic* set. Values for GFS initialization differ slightly from Table 2 because of the smaller sample. Leads: 24, 120, 240 h (2-m temperature, wind); days 1, 5, 9 ($T_{\max}$, $T_{\min}$); days 1, 5, 10 (precipitation).

Model	short	medium	long
<i>2-m temperature</i>			
GraphCast Op.	8.4 $\rightarrow$ 18.8	11.1 $\rightarrow$ 19.9	8.9 $\rightarrow$ 13.2
Pangu-Weather	8.5 $\rightarrow$ 17.1	11.8 $\rightarrow$ 18.2	4.0 $\rightarrow$ 7.3
FCN v2-small	6.0 $\rightarrow$ 8.1	7.1 $\rightarrow$ 9.2	5.5 $\rightarrow$ 7.1
<i>Daily $T_{\max}$</i>			
GraphCast Op.	7.9 $\rightarrow$ 19.0	8.8 $\rightarrow$ 19.7	7.1 $\rightarrow$ 13.4
Pangu-Weather	6.4 $\rightarrow$ 13.5	7.8 $\rightarrow$ 13.3	0.6 $\rightarrow$ 4.4
FCN v2-small	2.3 $\rightarrow$ 4.9	5.2 $\rightarrow$ 7.9	4.3 $\rightarrow$ 6.1
<i>Daily $T_{\min}$</i>			
GraphCast Op.	9.4 $\rightarrow$ 18.2	12.1 $\rightarrow$ 18.6	9.1 $\rightarrow$ 12.6
Pangu-Weather	15.5 $\rightarrow$ 20.2	14.2 $\rightarrow$ 18.4	5.3 $\rightarrow$ 8.2
FCN v2-small	11.5 $\rightarrow$ 13.2	9.1 $\rightarrow$ 10.9	4.9 $\rightarrow$ 7.3
<i>10-m wind speed</i>			
GraphCast Op.	0.7 $\rightarrow$ 3.1	1.8 $\rightarrow$ 5.9	0.8 $\rightarrow$ 2.2
Pangu-Weather	-1.8 $\rightarrow$ -0.7	1.4 $\rightarrow$ 3.4	0.2 $\rightarrow$ 0.7
FCN v2-small	-6.4 $\rightarrow$ -5.7	-1.7 $\rightarrow$ -0.4	0.1 $\rightarrow$ 0.4
<i>24-h precipitation</i>			
GraphCast Op.	4.8 $\rightarrow$ 9.4	6.2 $\rightarrow$ 11.7	2.7 $\rightarrow$ 5.8

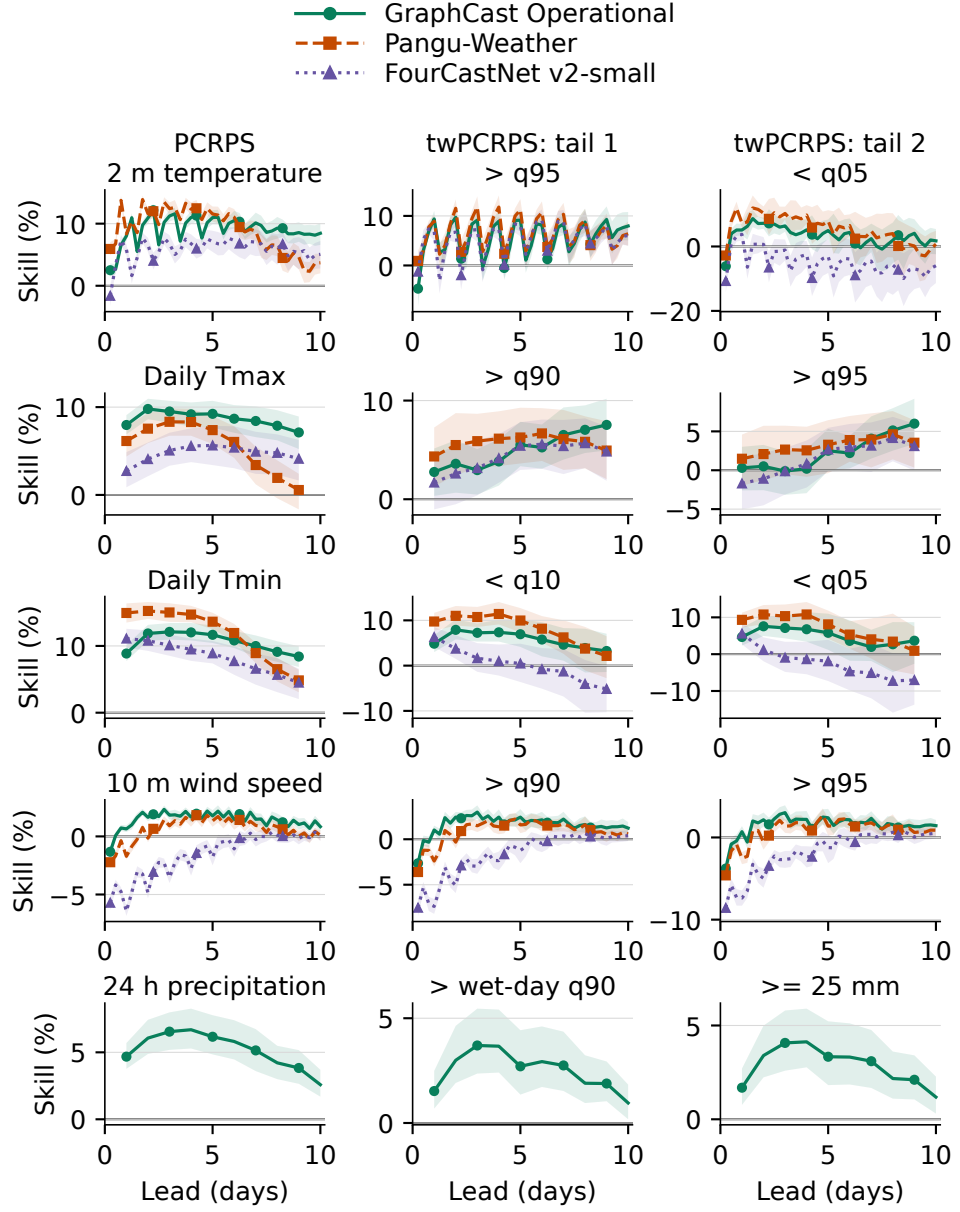


Fig. 2 Skill relative to GFS as a function of lead time. Rows are the five variables; columns show PCRPS, twPCRPS for tail 1 and tail 2. GraphCast, Pangu-Weather and FourCastNet v2-small use circles with solid lines, squares with dashed lines and upward triangles with dotted lines, respectively. Shading: 95% moving-block bootstrap confidence intervals.

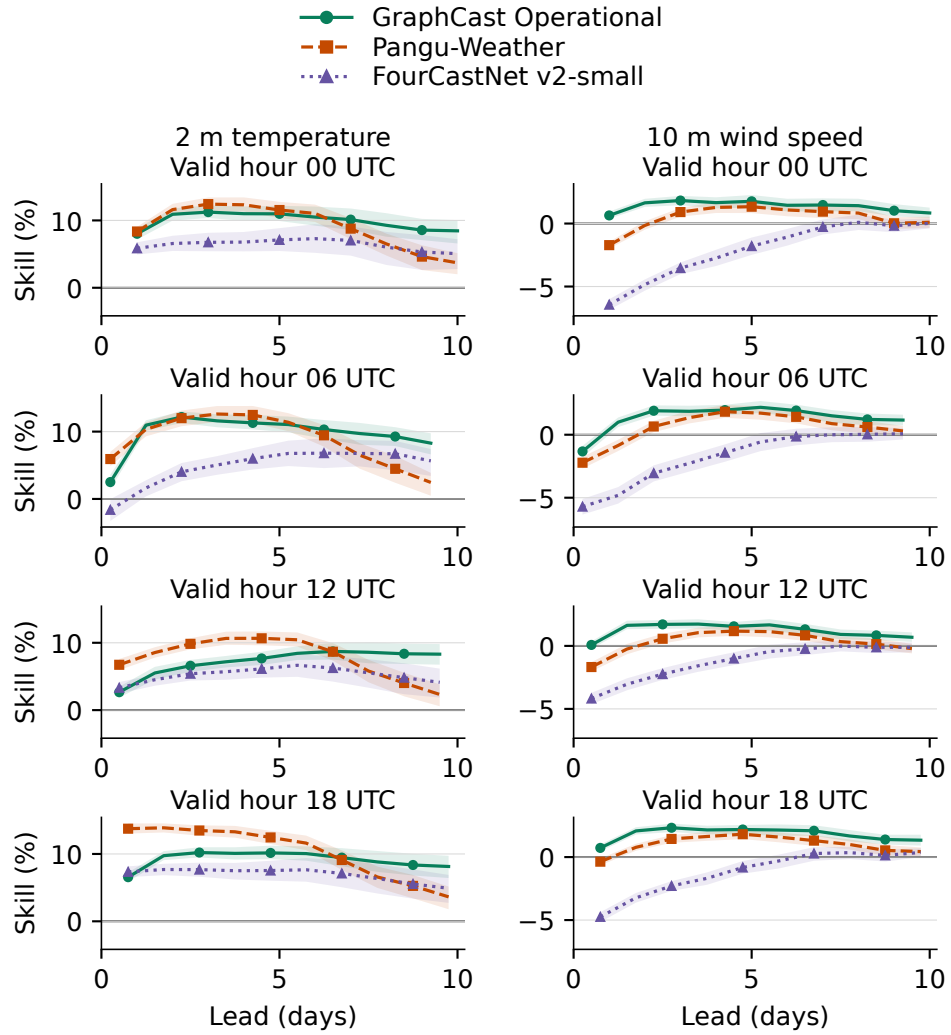


Fig. 3 PCRPS skill at every six-hourly lead, separated by valid UTC hour (rows), for 2-m temperature and 10-m wind speed (columns). Models use the line styles and marker shapes in Fig. 2. Shading gives 95% calendar-block bootstrap intervals.

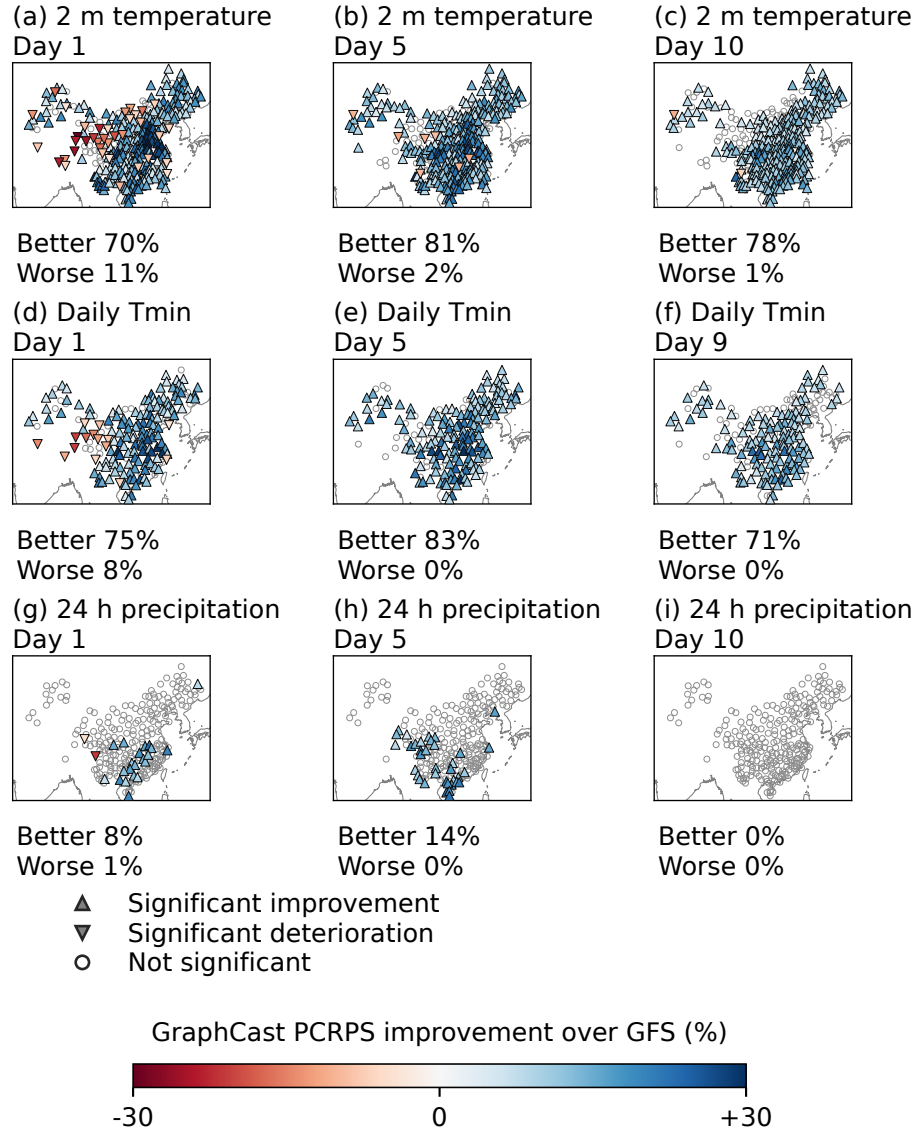


Fig. 4 Station-wise PCRPS skill of GraphCast relative to GFS for 2-m temperature, daily $T_{\min}$ and 24-h precipitation (rows) at three lead times (columns). Filled upward and downward triangles indicate significant improvement and deterioration, respectively, under the Diebold–Mariano test with Benjamini–Hochberg false discovery rate control at 5%; hollow circles indicate non-significant differences. Colour gives improvement magnitude. Better/worse percentages use all eligible stations in each panel.

2 m temperature Day 5				Daily Tmax Day 5			
Tropical (6)	+17*	+14*	+1	n=3	+21*	+20*	+13*
Temperate (149)	+14*	+17*	+10*	n=64	+17*	+17*	+14*
Arid (76)	+10*	+8*	+5*	n=54	+4*	-2	-1
Cold (75)	+11*	+7*	+5*	n=41	+9*	+3	+3
Highland (28)	+3	+14*	+8*	n=18	-7*	+7*	+0
All (334)	+11*	+12*	+7*	n=180	+9*	+7*	+6*
	GC	PW	FCN2		GC	PW	FCN2

Daily Tmin Day 5				10 m wind speed Day 5			
Tropical (3)	+12*	+11*	-1	n=6	+7*	+6*	+3*
Temperate (64)	+16*	+18*	+12*	n=149	+3*	+3*	-1*
Arid (54)	+11*	+12*	+8*	n=76	+2*	+1	-3*
Cold (41)	+11*	+10*	+5*	n=75	-1	-1*	-3*
Highland (18)	+5*	+16*	+12*	n=28	+1*	+1*	-1*
All (180)	+12*	+14*	+9*	n=334	+2*	+1*	-2*
	GC	PW	FCN2		GC	PW	FCN2

24 h precipitation Day 5		PCRPS skill vs GFS (%)	
Tropical (4)	+18*	GC: GraphCast	
Temperate (124)	+9*	PW: Pangu-Weather	
Arid (47)	+1*	FCN2: FourCastNet v2-small	
Cold (50)	+2*	+ / - : skill sign	
Highland (24)	+6*	* : 95% interval excludes 0	
All (249)	+6*	Counts: eligible stations	
	GC		

Fig. 5 PCRPS skill relative to GFS by climate group at day 5 (120 h for instantaneous variables). Signed cell values are skill percentages; asterisks indicate that the 95% bootstrap interval excludes zero. Rows follow Tropical, Temperate, Arid, Cold, Highland and All; eligible station counts are shown for each variable. GC: GraphCast; PW: Pangu-Weather; FCN2: FourCastNet v2-small.

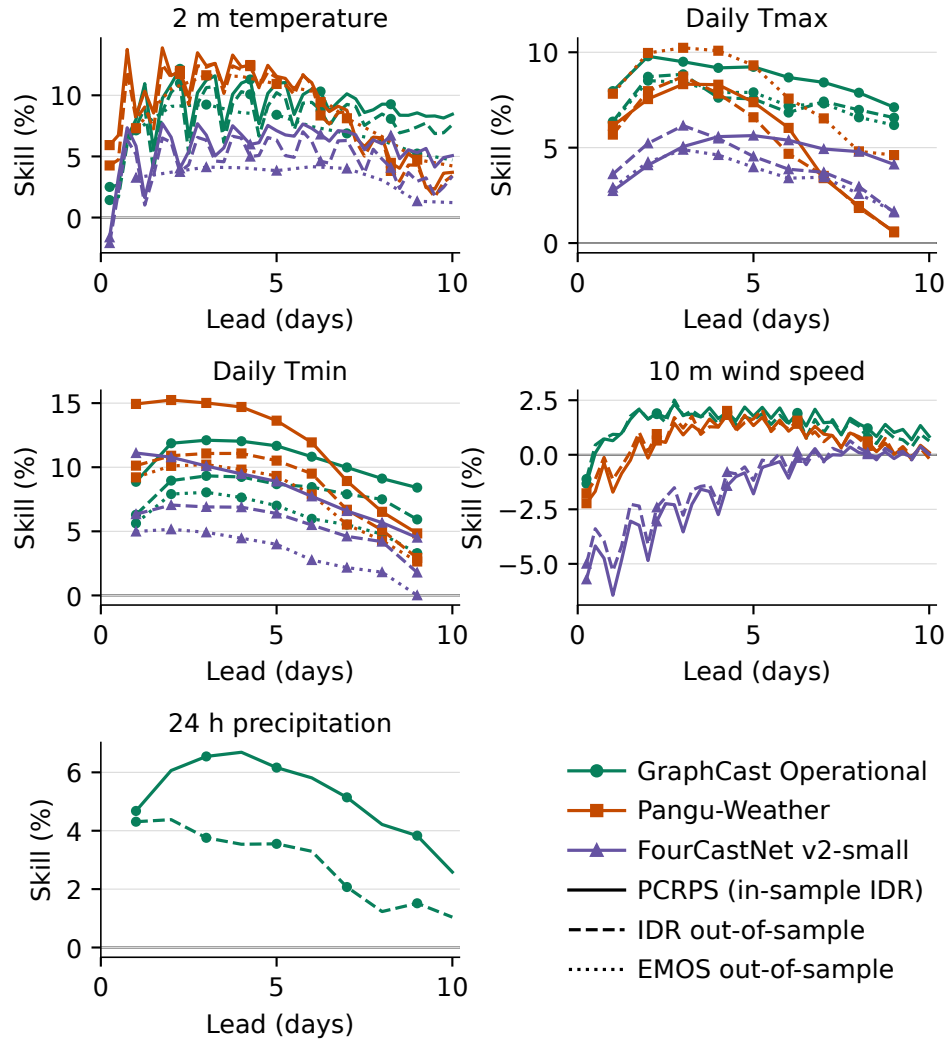


Fig. 6 In-sample PCRPS skill versus out-of-sample CRPS skill (IDR and, for temperature variables, EMOS; test period January 2024–August 2025) relative to GFS for the five variables. Circles, squares and upward triangles identify GraphCast, Pangu-Weather and FourCastNet v2-small. Solid, dashed and dotted lines indicate in-sample IDR, out-of-sample IDR and out-of-sample EMOS, respectively.

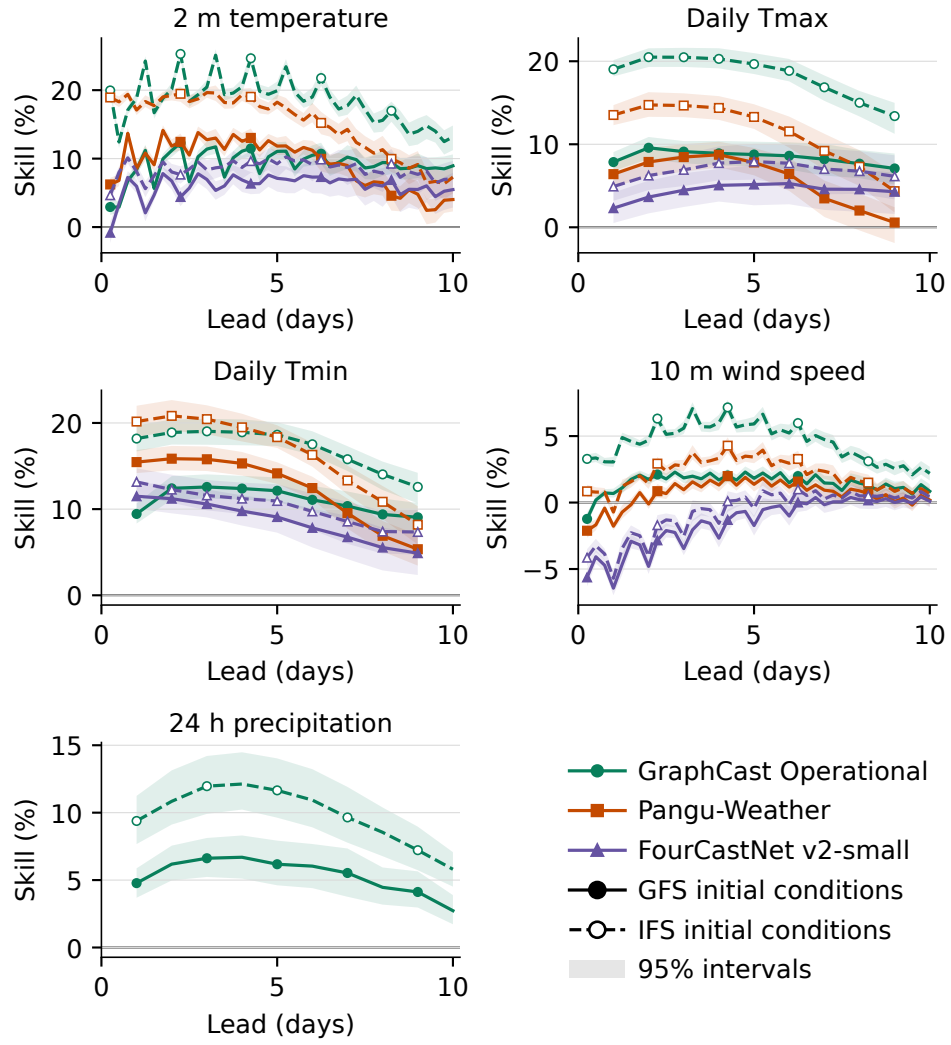


Fig. 7 PCRPS skill relative to GFS of GFS-initialized and IFS-initialized versions of each AI model, with 95% block-bootstrap confidence intervals (*ic* set). Marker shapes identify models as in Fig. 6; solid lines with filled markers and dashed lines with open markers indicate GFS and IFS initialization, respectively.

312 5 Discussion

313 *Extremes*

314 Using station observations rather than reanalysis, our results agree with Biegert
315 et al. (2026) and Olivetti and Messori (2024). When forecasts are compared through
316 calibrated predictive distributions and threshold-weighted proper scores, the in-
317 sample comparisons often favour AI forecasts for temperature tails. Strong-wind
318 gains are marginal, and the heavy-precipitation advantage weakens substantially
319 under independent-period calibration. Potential information should therefore be dis-
320 tinguished from realizable skill in warnings or deployed post-processing. The tail skill
321 is usually smaller than the skill for the full distribution, and it is not uniform. Four-
322 CastNet v2-small loses its advantage for cold extremes at long leads, and gains for
323 strong wind are marginal. Our design does not address record-breaking events, the
324 focus of Zhang et al. (2026), which are too rare in a four-year station sample for reli-
325 able inference. However, our tails are defined from an independent climatology and
326 scored without conditioning on the outcome, so they avoid the forecaster’s dilemma
327 (Lerch et al. 2017).

328 *Initial conditions*

329 The paired initialization comparison shows that skill depends strongly on the initial-
330 data source, especially at short leads. A mixed-source AI–GFS comparison does
331 not isolate architecture. The GFS-initialized comparison instead estimates relative
332 forecast-system skill conditional on GFS inputs; it is not necessarily a lower bound
333 for all AI applications. The weaker dependence of FourCastNet v2-small on the initial
334 conditions is consistent with its lower overall skill. We do not have a free station-level
335 baseline for IFS HRES, so we cannot say whether IFS-initialized AI models also beat
336 the IFS itself at these stations.

Regional variation

The reduced or negative skill at Highland and Arid stations, especially for $T_{\max}$, may reflect imperfect representation of near-surface temperature over complex terrain in the ERA5 training data of the AI models. It may also reflect a stronger diurnal cycle there that six-hourly sampling captures poorly. Testing these hypotheses is beyond the scope of this study. PCRPS removes constant offsets and strictly monotone recalibration errors, so regional differences are not explained by a constant bias alone. Non-monotone sampling and representativeness errors can still affect discrimination. Regional post-processing or downscaling should therefore not assume that AI models are uniformly superior.

Limitations

(i) Station series end in late August 2025, and the IFS-initialized runs have substantial gaps in 2024, which reduce the *ic* sample. (ii) Daily $T_{\max}/T_{\min}$ forecasts are derived from six-hourly output for all models, which underestimates the diurnal range; sub-daily extremes may change the comparison. (iii) Only GraphCast provides precipitation, so precipitation results cannot be generalised to other AI models. (iv) GFS was upgraded to version 16 on 22 March 2021; the *core* period (2022 onwards) uses a single major GFS version, but the *core21* set includes the transition. (v) The Tropical group is small, and the station set includes a few stations outside the mainland, whose influence has not been isolated in a dedicated geographic sensitivity test. (vi) PCRPS measures potential skill after optimal calibration. Raw forecasts, particularly for $T_{\max}$ and wind, have biases that would need post-processing in practice; the out-of-sample results support some overall temperature gains, but not universal rankings or uniformly positive extreme-tail skill.

361 6 Conclusions

362 We compared three GFS-initialized AI weather models with the operational GFS
363 at 334 stations in China over 2022–2025, using potential and threshold-weighted
364 scores, independent-period calibration, calendar-block resampling and station-wise
365 multiplicity-controlled tests. The main conclusions are:

- 366 1. Conditional on a shared GFS F000 input source, the comparison indicates positive
367 potential skill for the AI models at the reported daily temperature leads. Six-hourly
368 exceptions prevent a claim of superiority at every lead.
- 369 2. Temperature-tail improvements are more consistent than strong-wind gains.
370 GraphCast’s in-sample heavy-precipitation advantage does not establish reliable
371 out-of-sample tail superiority: out of sample, wet-day-q90 tail skill is significant
372 only at day 1.
- 373 3. Skill differs across climate groups and variables. Temperate stations generally show
374 positive gains, while GraphCast’s $T_{\max}$ PCRPS skill reverses sign at Highland
375 stations.
- 376 4. Switching the initialization source can substantially change short-range skill. This is
377 a conditional sensitivity of the trained forecast system, and fair comparisons should
378 report model versions, initialization products and calibration design explicitly.

379 Statements and Declarations

380 Funding

381 This research did not receive any specific grant from funding agencies in the public,
382 commercial, or not-for-profit sectors.

383 Competing interests

384 The author declares no competing interests.

385 **Author contributions**

386 **Yi Zou:** Conceptualization, Methodology, Software, Formal analysis, Investigation,
387 Data curation, Visualization, Writing – original draft, Writing – review & editing.

388 **Data and code availability**

389 AI model forecasts are available from the NOAA Open Data Dissemina-
390 tion program (<https://registry.opendata.aws/aiwp/>), GFS forecasts from [https://](https://registry.opendata.aws/noaa-gfs-bdp-pds/)
391 registry.opendata.aws/noaa-gfs-bdp-pds/, and ISD and GHCN-Daily observations
392 from NOAA’s National Centers for Environmental Information. The Köppen-Geiger
393 maps are described by Beck et al. (2023). Code, regression tests, station lists, full
394 machine-readable derived result tables, method metadata and run/provenance records
395 are publicly available in the GitHub repository at [https://github.com/yizou728282/](https://github.com/yizou728282/nwp-ai-extremes-china-obs-public)
396 `nwp-ai-extremes-china-obs-public`, release `submission-2026-10-04`. The release
397 excludes the large forecast-point and observation caches, auxiliary map datasets, anal-
398 ysis cubes and case-level score arrays; these can be regenerated from the public sources
399 using the supplied scripts. Online Resource 1 (`ESM_1.pdf`) contains supplementary
400 key-result tables, the station inventory and links to the repository and release.

401 **Supplementary Information**

402 Online Resource 1 (`ESM_1.pdf`). Supplementary key-result tables for the core, core21
403 and paired-initialization comparisons, station inventory, and links to the public code
404 and derived-data repository. Full machine-readable results, metadata and provenance
405 records are available in the linked repository.

406 **Acknowledgements**

407 I thank NOAA and CIRA for making the AIWP archive publicly available, and NOAA
408 NCEI for the ISD and GHCN-Daily datasets.

References

- Allen S, Ginsbourger D, Ziegel J (2023) Evaluating Forecasts for High-Impact Events Using Transformed Kernel Scores. *SIAM/ASA Journal on Uncertainty Quantification* 11(3):906–940. <https://doi.org/10.1137/22M1532184>
- Beck HE, McVicar TR, Vergopolan N, Berg A, Lutsko NJ, Dufour A, Zeng Z, Jiang X, van Dijk AIJM, Miralles DG (2023) High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections. *Scientific Data* 10(1):724. <https://doi.org/10.1038/s41597-023-02549-6>
- Ben Bouallègue Z, Clare MCA, Magnusson L, Gascón E, Maier-Gerber M, Janoušek M, Rodwell M, Pinault F, Dramsch JS, Lang STK, Raoult B, Rabier F, Chevallier M, Sandu I, Dueben P, Chantry M, Pappenberger F (2024) The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society* 105(6):E864–E883. <https://doi.org/10.1175/BAMS-D-23-0162.1>
- Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q (2023) Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 619(7970):533–538. <https://doi.org/10.1038/s41586-023-06185-3>
- Biegert T, Allen S, Alber A, Lerch S (2026) Towards fair comparisons of AI- and physics-based weather models for extreme events via the weighted potential CRPS.

433 arXiv preprint arXiv:2606.21170, <https://doi.org/10.48550/arXiv.2606.21170>

434 Bonev B, Kurth T, Hundt C, Pathak J, Baust M, Kashinath K, Anandkumar
435 A (2023) Spherical Fourier neural operators: Learning stable dynamics on the
436 sphere. In: Proceedings of the 40th International Conference on Machine Learning,
437 arXiv:2306.03838

438 Charlton-Perez AJ, Dacre HF, Driscoll S, Gray SL, Harvey B, Harvey NJ, Hunt KMR,
439 Lee RW, Swaminathan R, Vandaele R, Volonté A (2024) Do AI models produce
440 better weather forecasts than physics-based models? A quantitative evaluation case
441 study of Storm Ciarán. *npj Climate and Atmospheric Science* 7(1):93. [https://](https://doi.org/10.1038/s41612-024-00638-w)
442 doi.org/10.1038/s41612-024-00638-w

443 Diebold FX, Mariano RS (1995) Comparing Predictive Accuracy. *Journal*
444 *of Business & Economic Statistics* 13(3):253–263. [https://doi.org/10.1080/](https://doi.org/10.1080/07350015.1995.10524599)
445 [07350015.1995.10524599](https://doi.org/10.1080/07350015.1995.10524599)

446 Gneiting T, Raftery AE (2007) Strictly Proper Scoring Rules, Prediction, and Esti-
447 mation. *Journal of the American Statistical Association* 102(477):359–378. [https://](https://doi.org/10.1198/016214506000001437)
448 doi.org/10.1198/016214506000001437

449 Gneiting T, Ranjan R (2011) Comparing Density Forecasts Using Threshold- and
450 Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*
451 29(3):411–422. <https://doi.org/10.1198/jbes.2010.08110>

452 Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated Probabilis-
453 tic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS
454 Estimation. *Monthly Weather Review* 133(5):1098–1118. [https://doi.org/10.1175/](https://doi.org/10.1175/MWR2904.1)
455 [MWR2904.1](https://doi.org/10.1175/MWR2904.1)

- 456 Gneiting T, Biegert T, Kraus K, Walz EM, Jordan AI, Lerch S (2026) Probabilistic
457 Measures Afford Fair Comparisons of AIWP and NWP Model Output. *Artificial
458 Intelligence for the Earth Systems* 5(1). <https://doi.org/10.1175/AIES-D-25-0054.1>
- 459 He P, Jin C, Dong Y, Qi C, Lei T (2026) Machine Learning-Based Postprocessing
460 Ensemble of AI and NWP Models for Short-Term Air Temperature Forecasting
461 Over Beijing. *Meteorological Applications* 33(5):e70238. [https://doi.org/10.1002/
462 met.70238](https://doi.org/10.1002/met.70238)
- 463 Henzi A, Ziegel JF, Gneiting T (2021) Isotonic Distributional Regression. *Journal
464 of the Royal Statistical Society Series B: Statistical Methodology* 83(5):963–993.
465 <https://doi.org/10.1111/rssb.12450>
- 466 Kunsch HR (1989) The Jackknife and the Bootstrap for General Stationary Obser-
467 vations. *The Annals of Statistics* 17(3):1217–1241. [https://doi.org/10.1214/aos/
468 1176347265](https://doi.org/10.1214/aos/1176347265)
- 469 Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger P, Fortunato M, Alet F, Ravuri
470 S, Ewalds T, Eaton-Rosen Z, Hu W, Merose A, Hoyer S, Holland G, Vinyals O,
471 Stott J, Pritzel A, Mohamed S, Battaglia P (2023) Learning skillful medium-range
472 global weather forecasting. *Science* 382(6677):1416–1421. [https://doi.org/10.1126/
473 science.adi2336](https://doi.org/10.1126/science.adi2336)
- 474 Lerch S, Thorarindottir TL, Ravazzolo F, Gneiting T (2017) Forecaster’s Dilemma:
475 Extreme Events and Forecast Evaluation. *Statistical Science* 32(1):106–127. <https://doi.org/10.1214/16-STS588>
476
- 477 Matheson JE, Winkler RL (1976) Scoring Rules for Continuous Probability
478 Distributions. *Management Science* 22(10):1087–1096. [https://doi.org/10.1287/
479 mnsc.22.10.1087](https://doi.org/10.1287/mnsc.22.10.1087)

- 480 Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG (2012) An Overview of
481 the Global Historical Climatology Network-Daily Database. *Journal of Atmo-*
482 *spheric and Oceanic Technology* 29(7):897–910. <https://doi.org/10.1175/JTECH->
483 [D-11-00103.1](https://doi.org/10.1175/JTECH-D-11-00103.1)
- 484 Newey WK, West KD (1987) A Simple, Positive Semi-Definite, Heteroskedasticity
485 and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55(3):703–708.
486 <https://doi.org/10.2307/1913610>
- 487 NOAA National Centers for Environmental Information (2018) Integrated Surface
488 Data (ISD): Data Documentation. URL [https://www.ncei.noaa.gov/pub/data/](https://www.ncei.noaa.gov/pub/data/noaa/isd-format-document.pdf)
489 [noaa/isd-format-document.pdf](https://www.ncei.noaa.gov/pub/data/noaa/isd-format-document.pdf), january 12, 2018; liquid-precipitation AA1–AA4
490 fields
- 491 Olivetti L, Messori G (2024) Do data-driven models beat numerical models in
492 forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and
493 GraphCast. *Geoscientific Model Development* 17(21):7915–7962. [https://doi.org/](https://doi.org/10.5194/gmd-17-7915-2024)
494 [10.5194/gmd-17-7915-2024](https://doi.org/10.5194/gmd-17-7915-2024)
- 495 Radford JT, Ebert-Uphoff I, Stewart JQ, Musgrave KD, DeMaria R, Tourville N,
496 Hilburn K (2025) Accelerating Community-Wide Evaluation of AI Models for
497 Global Weather Prediction by Facilitating Access to Model Output. *Bulletin of the*
498 *American Meteorological Society* 106(1):E68–E76. <https://doi.org/10.1175/BAMS->
499 [D-24-0057.1](https://doi.org/10.1175/BAMS-D-24-0057.1)
- 500 Rasp S, Hoyer S, Merose A, Langmore I, Battaglia P, Russell T, Sanchez-Gonzalez
501 A, Yang V, Carver R, Agrawal S, Chantry M, Ben Bouallegue Z, Dueben P,
502 Bromberg C, Sisk J, Barrington L, Bell A, Sha F (2024) WeatherBench 2: A Bench-
503 mark for the Next Generation of Data-Driven Global Weather Models. *Journal*
504 *of Advances in Modeling Earth Systems* 16(6):e2023MS004019. <https://doi.org/>

505 10.1029/2023MS004019

506 Smith A, Lott N, Vose R (2011) The Integrated Surface Database: Recent Devel-
507 opments and Partnerships. *Bulletin of the American Meteorological Society*
508 92(6):704–708. <https://doi.org/10.1175/2011BAMS3015.1>

509 Walz EM, Henzi A, Ziegel J, Gneiting T (2024) Easy Uncertainty Quantification
510 (EasyUQ): Generating Predictive Distributions from Single-Valued Model Output.
511 *SIAM Review* 66(1):91–122. <https://doi.org/10.1137/22M1541915>

512 Wilks DS (2016) “The Stippling Shows Statistically Significant Grid Points”: How
513 Research Results are Routinely Overstated and Overinterpreted, and What to Do
514 about It. *Bulletin of the American Meteorological Society* 97(12):2263–2273. <https://doi.org/10.1175/BAMS-D-15-00267.1>

516 World Meteorological Organization (1995) Manual on Codes, Volume I.1: Inter-
517 national Codes, Part A. URL [https://mgm.gov.tr/FILES/resmi-istatistikler/](https://mgm.gov.tr/FILES/resmi-istatistikler/WMO_NO_306.pdf)
518 WMO_NO_306.pdf, regulation 12.4.9, 24-hour precipitation in SYNOP section 3

519 Yan Z, Lu X, Wu L, Liu F, Qiu R, Cui Y, Ma X (2025) Evaluation of precipitation
520 forecasting base on GraphCast over mainland China. *Scientific Reports* 15:14771.
521 <https://doi.org/10.1038/s41598-025-98944-7>

522 Zhang Z, Fischer E, Zscheischler J, Engelke S (2026) Physics-based models out-
523 perform AI weather forecasts of record-breaking extremes. *Science Advances*
524 12(18):eac1433. <https://doi.org/10.1126/sciadv.aec1433>